

DATA HANDLING PART I



SCIANTA ANALYTICS
DEEP INSIGHT™

“The natural evolution of machine learning, Cognitive Computing attempts to imbue, in computer systems, the same insight and understanding we see in humans.”

Earl Cox
Chief Scientist, Scianta Analytics
Splunk .Conf 2013



SCIENTA ANALYTICS
DEEP INSIGHT™

©2014-2018 Scianta Analytics LLC, All Rights Reserved

AGENDA

Introduction to Machine Intelligence	Data Handling 1	Data Handling 2	Anomaly Detection	Transactional Behavior	Impact Analysis
Academic Concepts	Collection	Retention	Anomaly Definition	Defining Transactions	Organizational Visibility
Data Systems	Storage	Format	Measuring Normality	Transaction Relationships	Types of Impact
Maturity Curve	Security	Labeling		Probability Measurement	Responsiveness

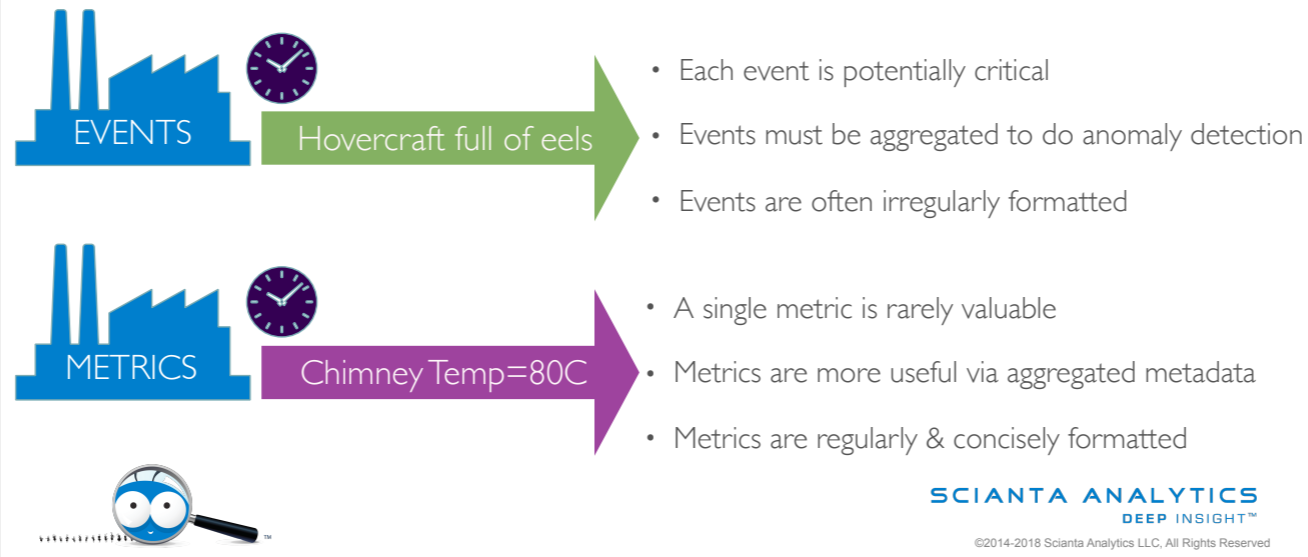


SCIANTA ANALYTICS
DEEP INSIGHT™

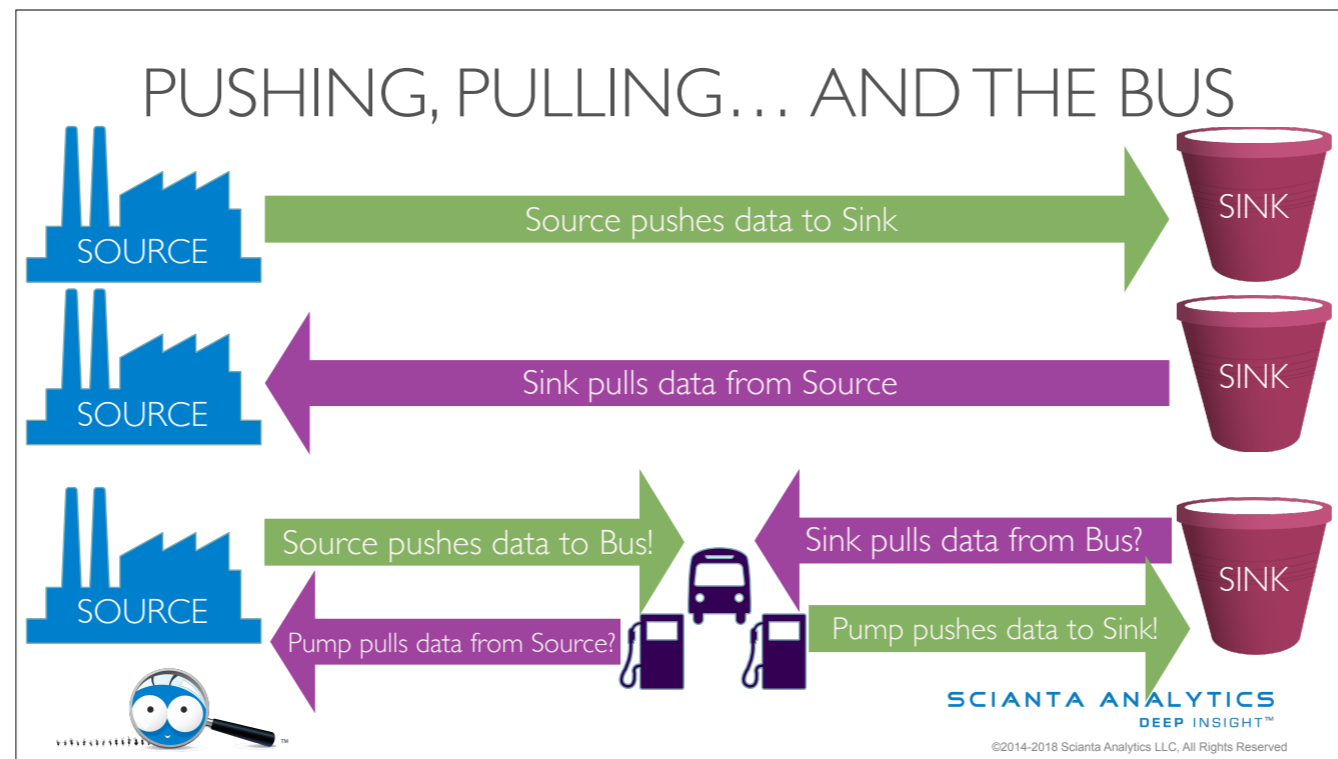
©2014-2018 Scianta Analytics LLC, All Rights Reserved

DATA COLLECTION

Events and Metrics



Because these two types of data have different characteristics, it's sometimes useful to store them differently. However, there are caveats: the benefit of easy time synchronization and the ability to use a single search tool for both types can outweigh a lot of storage inefficiency. Additionally, aggregating the data to find what's normal or abnormal is very valuable, and using a single tool for this has benefits.



Whether it's events or metrics, there's a few standard ways to get data from the source to the sink; note that there are dozens of implementations and confusing brand names, but everything fits one of these patterns.


- * One: your source pushes to your sink. Syslog is a common example of this, or you might see a logging agent like Splunk's forwarder or Logstash. This is a good design for big piles of event data, like logfiles. It doesn't work as well with data sources that need interaction, like an API or a database.
- * Two: for those types of data, the common pattern is to have your Sink pull from the source. This allows you to run complex code, but it doesn't scale well because you typically have fewer sinks than sources.
- * Three: the most complex pattern uses an abstraction layer called a message bus, often with extra code bits to handle pumping data around. AWS Kinesis and Apache Kafka are the most common examples of this pattern now. It's more complicated, but the bus scales well and provides fault tolerance.

UNDERSTANDING BARRIERS

No.

As a _____
I want to _____
so that I _____

Physical
Data Link
Network
Transport
Session
Presentation
Application
Money
Politics
Law
Fashion


SCIANTA ANALYTICS
 DEEP INSIGHT™
©2014-2018 Scianta Analytics LLC, All Rights Reserved

In a perfect world, scalability and data format would be your biggest worries around data handling... but it isn't a perfect world, and the hardest problem is probably access. Getting at the data required to answer a question isn't always about systems connectivity. You'll also need to consider who owns the data, what they think of your team's mission, what the cost-benefit scenario is for this data, whether it's legal to use the data as you want, and whether the tools you intend to use are going to be adopted by the people who need to use them.

The best advice we can give is to be open about your goals, ask a lot of questions, and be willing to share. You're a lot more likely to be successful if the data owner is benefitting as well as your team. Perhaps you're collecting data for a large initiative; can you give the data owner advance warning of detected problems in their specific area? Perhaps you're collecting data for security; can you give the data owner an operations dashboard as well? Not only do these efforts get your job done more quickly, they also help you uncover scenarios that you may not have considered on your own. No one is a subject matter expert in everything.

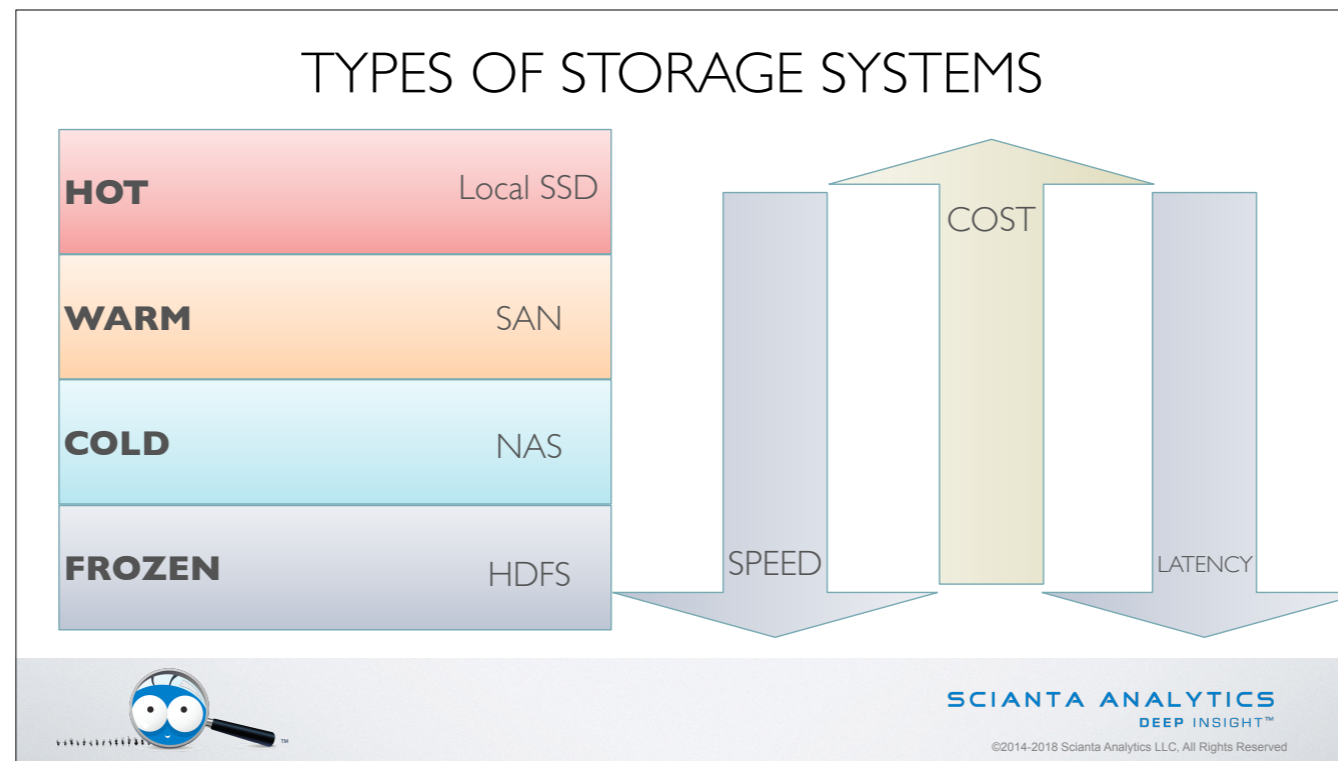
AGENDA

Introduction to Machine Intelligence	Data Handling 1	Data Handling 2	Anomaly Detection	Transactional Behavior	Impact Analysis
Academic Concepts	Collection	Retention	Anomaly Definition	Defining Transactions	Organizational Visibility
Data Systems	Storage	Format	Measuring Normality	Transaction Relationships	Types of Impact
Maturity Curve	Security	Labeling		Probability Measurement	Responsiveness



SCIANTA ANALYTICS
DEEP INSIGHT™

©2014-2018 Scianta Analytics LLC, All Rights Reserved



The first gotcha in this type of data processing is that your storage must be equally fast at reading and writing. Many storage systems optimize for one or the other, on the theory that writing to disk and reading from disk can be done by separate systems. This is a workable assumption for static systems processing well-understood data, but cognitive computing at scale is usually aimed at dynamically achieving understanding. Flexible cognitive computing and ad hoc search use cases require putting compute and storage close to each other, and cannot optimize pipelines for read-only or write-only. Instead, storage is typically structured in time-based tiers, allowing the best experience to be provided for the freshest data. An automated roll of data blocks down through the hierarchy allows storage administrators to provide less expensive and less performant storage options for older data.

Technology changes rapidly and different organizations have different tolerances for cost, particularly when infrastructural services are involved. Rather than focus on particular technologies, we'll advise the following: work closely with the vendor of your data processing platform and your storage provider, get detailed about the concurrent read/write performance for large and small blocks of data, and keep an eye on latency. The recommendations on this slide are just suggestions, not requirements.

COLUMNAR ACCELERATION

- `<log><time>16:33:50 Feb 4, 2018</time><fwlog><user>alice</user><message>VPN login denied, bad cred</message></log>`
- `16:35:20 4 February 2018 fwlog VPN login denied for bob, 503.44`
- `4:37:14 PM 02-04-18 host=fwlog service=VPN action=login result=denied user=cindy reason=credfail`
- `[{"2018-02-04T16:38:52.627Z","error",{"message":"VPN login fail","user":"dave","reason":"credential failure","host":"fwlog"}]}`

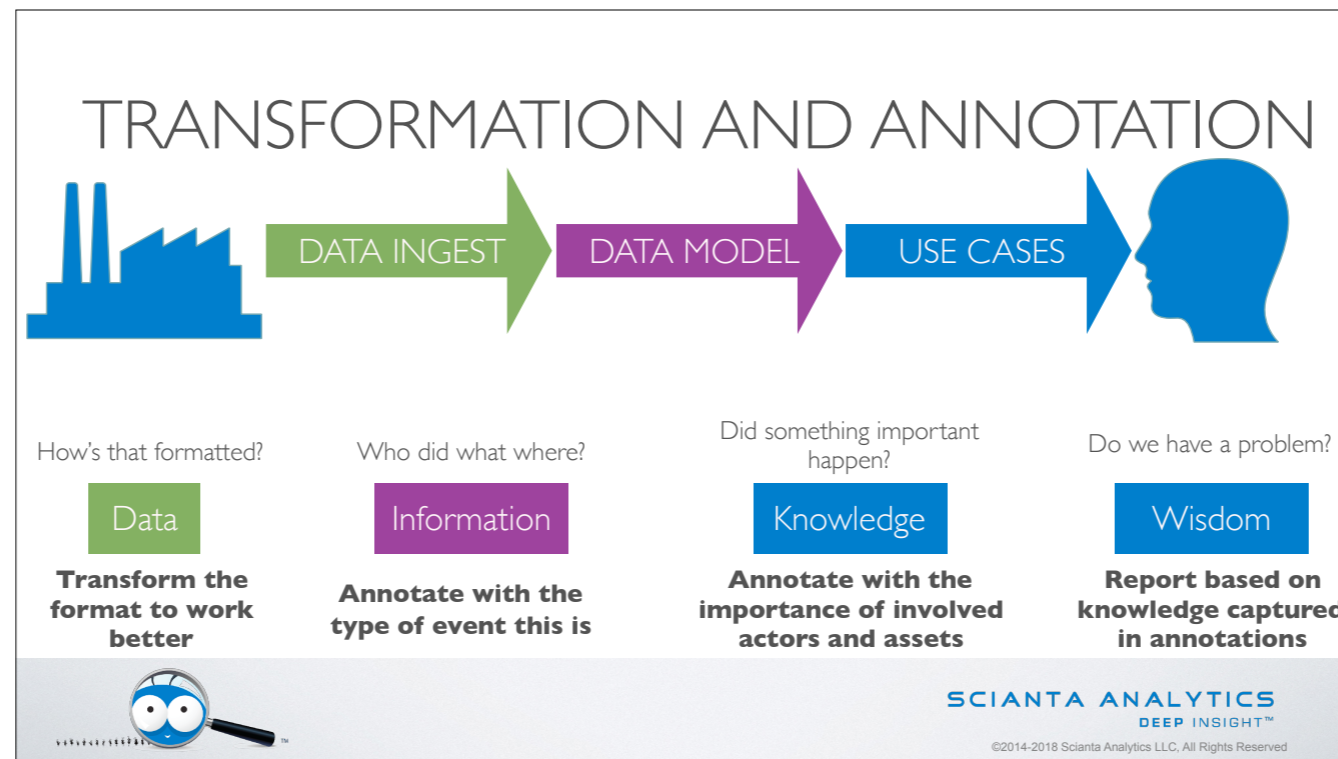
Time	Host	User	Action	Result
1517790830	fwlog	alice	vpn-login	denied
1517790920	fwlog	bob	vpn-login	denied
1517791034	fwlog	cindy	vpn-login	denied
1517791132	fwlog	dave	vpn-login	denied



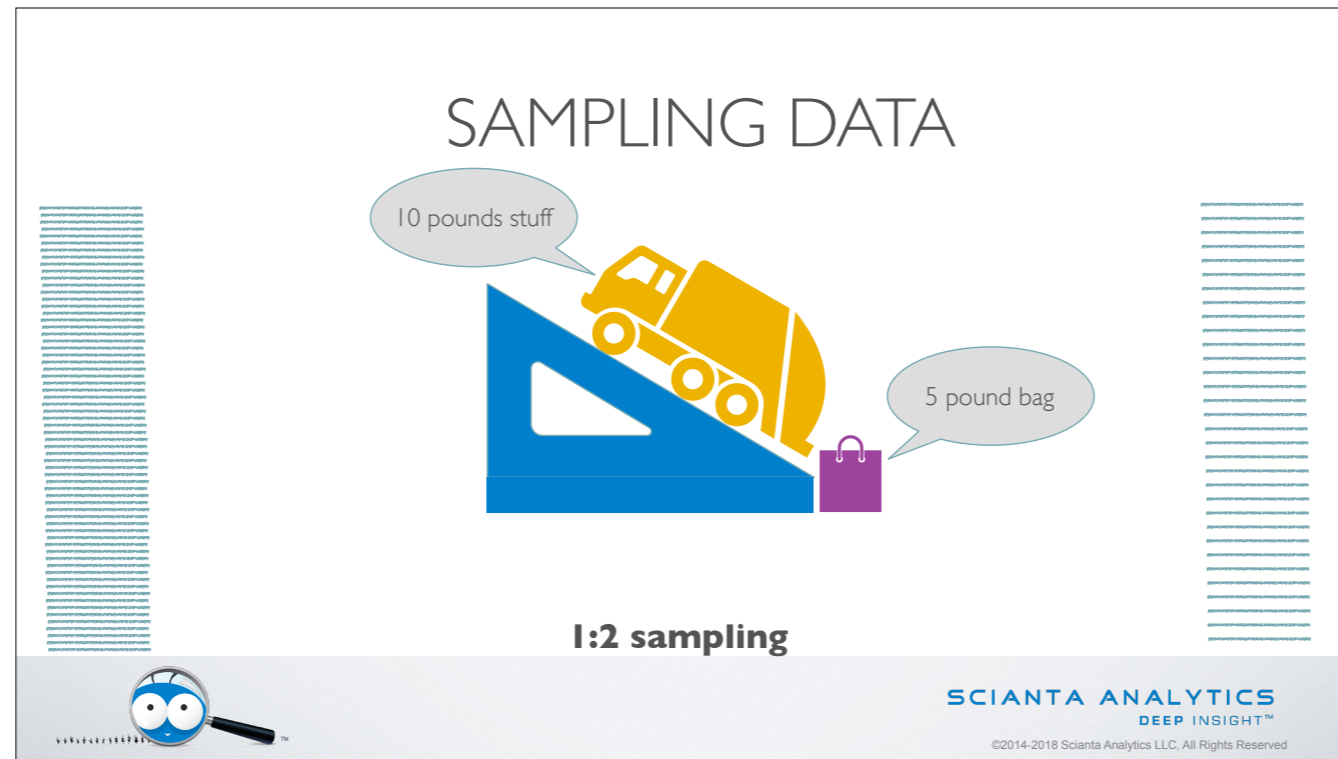
SCIANTA ANALYTICS
DEEP INSIGHT™

©2014-2018 Scianta Analytics LLC, All Rights Reserved

Remember in Events and Metrics when we talked about aggregating data for analysis? Columnar Acceleration is a popular and powerful way to do that. This is the process of converting specific events in their raw formats into semantically legible events with a shared format. Searching a columnar store is blazingly fast when compared to searching raw data. However, converting the data from raw formats to columnar formats takes processing power. Like everything in performance tuning, there's an element of borrowing from Peter to pay Paul. Still, columnar acceleration is so useful for large amounts of data that the real question is WHEN to apply it, not IF. Generally, the earlier in the data collection pipeline that data can be accelerated, the better.



That provides an excellent segue to the concepts of transforming and annotating data, or in other words labelling our data. Again, the earlier in the pipeline that these actions can be taken the better, but it's not always possible until a human has reviewed. Our mission is to understand.



We've talked about the cost of storage and we've talked about the value of transforming data in the pipeline... So it stands to reason, as long as you're transforming in that pipeline, how about reducing storage cost by throwing away some redundant data?

This can be a useful approach in some scenarios, but caution is required. Let's start with the easy story, Metrics.

Metrics are the ideal data for statistical analysis because outliers are typically really that; measurement errors or weird flukes that can be ignored. The further back in time one goes, the less important a given measurement becomes; aggregates over longer and longer time windows are good enough.

Sampling metrics at ingest and at movement between storage layers is a fairly normal procedure.

With Events, the story is more complicated. A single event can be the critical piece of information explaining what went wrong. Still, strange events are not helpful when you're trying to train an algorithm to recognize what's normal. Sampling can help reduce the noise in training data, which is overall beneficial. Just don't sample your raw event data; you don't want to lose events or fail to test them against your cognitive computing models if you can help it.

AGENDA

Introduction to Machine Intelligence	Data Handling 1	Data Handling 2	Anomaly Detection	Transactional Behavior	Impact Analysis
Academic Concepts	Collection	Retention	Anomaly Definition	Defining Transactions	Organizational Visibility
Data Systems	Storage	Format	Measuring Normality	Transaction Relationships	Types of Impact
Maturity Curve	Security	Labeling		Probability Measurement	Responsiveness



SCIANTA ANALYTICS
DEEP INSIGHT™

©2014-2018 Scianta Analytics LLC, All Rights Reserved



Let's say you're wildly successful with your cognitive computing system, you've got the data pulse of the organization, and your middle managers and executives are pulling useful wisdom from it. Congratulations, you've built a high value target! Your system is now a tempting attack point for bad people investigating your organization. Maybe the goal is industrial espionage, maybe it's financial gain, but there's a lot of data being converted to useful forms and a lot of powerful user accounts coming here.

All is not lost, though, we'll touch on security posture mitigations shortly.

REGULATORY CONCERNS

Data Governance

User Privacy



Financial Due Diligence

Ethical Management



SCIANTA ANALYTICS
DEEP INSIGHT™

©2014-2018 Scianta Analytics LLC, All Rights Reserved

It's a common truism that security and compliance aren't the same thing at all, but the same teams are often in charge of assuring both. Compliance testing can be a useful baseline for securing a system. And more importantly, failing to handle compliance properly is easier to detect and more certain to produce a costly outcome than failing to handle security properly. The wily hacker might come by, but the internal auditor absolutely will.



So after thinking through the threat landscape and regulatory landscape, you've probably got some specifics to watch for. You might operate in many different jurisdictions, or you might have access to credit cards or health insurance accounts. Still, there's some commonalities for anyone with a high value data target.

If you don't collect or retain the data, it can't be stolen.

If your users can't access data they're not supposed to see, then they can't lose that data.

If you patch your known vulnerabilities and disable unused services, you reduce the number of access paths.

And most importantly, use the cognitive computing tools at your disposal on your own log data! If you detect anomalies, unusual access patterns, and troubling data flows in your own data system, these are just as important as the ones you're providing to your users.

“The natural evolution of machine learning, Cognitive Computing attempts to imbue, in computer systems, the same insight and understanding we see in humans.”

Earl Cox
Chief Scientist, Scianta Analytics
Splunk .Conf 2013



SCIANTA ANALYTICS
DEEP INSIGHT™

©2014-2018 Scianta Analytics LLC, All Rights Reserved

Computers are force multipliers, not analysts; but they can help an analyst be more productive and successful. I hope this has been a useful overview of data handling; next, we will drill deeper into some of the techniques we've just covered. Thank you!



Thank you!